# Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/US04/023676

International filing date: 22 July 2004 (22.07.2004)

Document type: Certified copy of priority document

Document details: Country/Office: US
Number: 60/489,589
Filing date: 23 July 2003 (23.07.2003)

Date of receipt at the International Bureau: 19 August 2004 (19.08.2004)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)




World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

1210734

# THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

*August 12, 2004*

**THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE.**

**APPLICATION NUMBER:** *60/489,589*
**FILING DATE:** *July 23, 2003*
**RELATED PCT APPLICATION NUMBER:** *PCT/US04/23676*

Certified by

Jon W Dudas

Acting Under Secretary of Commerce
for Intellectual Property
and Acting Director of the U.S.
Patent and Trademark Office

SUBSTITUTE PTO/SB/16 (5-03)



# PROVISIONAL APPLICATION FOR PATENT COVER SHEET

**This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR §1.53(c).**

## INVENTOR(S)

| Given Name (first and middle (if any)) | Family Name or Surname | Residence (City and either State or Foreign Country) |
|---|---|---|
| Robert | Morris | Carrboro, NC |

Additional inventors are being named on the 0 separately numbered sheets attached hereto

## TITLE OF THE INVENTION (500 characters max)

Spoken Word Spotting Queries

## CORRESPONDENCE ADDRESS

*Direct all correspondence to:*

[X] Customer Number: 26161

***OR***

| | | | | | |
|---|---|---|---|---|---|
| [ ] Firm *or* Individual Name | | | | | |
| Address | | | | | |
| Address | | | | | |
| City | | State | | ZIP | |
| Country | United States | Telephone | | Fax | |

## ENCLOSED APPLICATION PARTS *(check all that apply)*

[X] Specification — *Number of Pages* 12

[X] Drawing(s) — *Number of Sheets* 2

[ ] Application Data Sheet. See 37 CFR 1.76.

[ ] CD(s), Number

[X] Other (specify) Appendix A (17 pages)

## METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

[X] Applicant Claims small entity status. See 37 CFR 1.27.

[X] A check or money order is enclosed to cover the filing fees.

[ ] The Director is hereby authorized to charge filing fees or credit any overpayment to Deposit Account Number: 06-1050

[ ] Payment by credit card. Form PTO-2038 is attached.

FILING FEE AMOUNT ($): $80

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

[X] No.

[ ] Yes, the name of the U.S. Government agency and the Government contract number are:

*Respectfully submitted,*

Signature Jan Robin Rohlicek

Date July 23, 2003

Typed Name J. Robin Rohlicek, J.D., Ph.D., Reg. No. 43,349

Telephone No. (617) 542-5070

Docket No. 14061-004P01

20696373.doc

**CERTIFICATE OF MAILING BY EXPRESS MAIL**

Express Mail Label No. EV303681663US

Date of Deposit July 23, 2003

Attorney's Docket No.: 14061-004P01

# PROVISIONAL APPLICATION FOR PATENT

## under

## 37 CFR §1.53(c)

TITLE: SPOKEN WORD SPOTTING QUERIES

APPLICANT: ROBERT MORRIS

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EV303681663US

July 23, 2003

Date of Deposit

# SPOKEN WORD SPOTTING QUERIES

## Background

[01] This invention relates to word spotting using spoken queries.

[02] Word spotting (which is understood to include phrase spotting, spotting of more complex linguistically-based events, and related techniques for detection of events) is a type of speech recognition in which occurrences of linguistically-based events are detected in an input acoustically-based signal. Word spotting, as well as speech recognition in general, has been performed using phonetically-based statistical models. In such word spotting systems, a query is represented in terms of phonetic units, for instance as a sequence of phonemes, which are then used to construct statistical models based on parameters associated with the phonemes.

[03] When a query is represented in text form, it can be converted into a phonetic representation using dictionaries and/or linguistic rules. The accuracy of the phonetic representation can affect the ability of the word spotting system to detect occurrences of the query.

## Summary

[04] In one aspect, in general, the invention features a method, and corresponding system and computer software, in which query data from one or more spoken instance of a query are accepted, and then processed. Processing the query data including determining a representation of the query that defines multiple sequences of subword units each representing the query. Then putative instances of the query are located in input data from an audio signal using the determined representation of the query.

[05] Aspects of the invention can include one or more of the following features:

[06] The query can represent a single word, a phrase or sequence of words, a larger linguistic unit, or a complex query such as a Boolean query of a query that includes components such as a wildcard or a time interval.

[07] A speech recognition algorithm, which may be implemented as a software procedure and/or a hardware circuit, is applied to the query data. The speech recognition

algorithm can be a statistical algorithm, such as one based on Hidden Markov Models (HMMs), or be based on other pattern matching approaches.

[08] The query data can represent the spoken instances of the query as waveform samples, signal processing features, or other acoustically-based data. The query data can also include the result of application of another speech recognition algorithm or procedure.

[09] The subword units can include linguistic units, such as phonetically-based units.

[010] A word spotting algorithm configured using the determined representation of the query can be applied to locate the putative instances of the query.

[011] Parameter values of the speech recognition algorithm for application to the query data are selected according to characteristics of the word spotting algorithm. The parameter values of the speech recognition algorithm can be selected to optimize an accuracy (or other performance measure) of the word spotting algorithm. For example, an expected detection rate or a false alarm rate or a combination of the two can be optimized.

[012] The parameters for which values can be selected can include one or more of an insertion factor, a recognition search beam width, a recognition grammar factor, and a number of recognition hypotheses.

[013] Determining the representation of the query can include determining a network of the subword units. The multiple sequences of subword units can then correspond to different paths through the network.

[014] An n-best list of recognition results can also be determined, and each of the multiple sequences of subword units can correspond to a different one in the n-best list of recognition results.

[015] Audio data representing the spoken utterances of the query spoken by a user, and processed to form the query data.

[016] A user can make a selection portions of stored data from a previously accepted audio signal, and these portions of the stored data are processed to form the query data.

[017] Prior to accepting the selection by the user, the previously accepted audio signal can be processed according to a first speech recognition algorithm to produce the stored data. This first speech recognition algorithm can produce data related to presence of the subword units at different times in the audio signal. Processing the query data then includes applying a second speech recognition algorithm to the query data.

[018] Aspects of the invention can include one or more of the following advantages.

[019] By choosing the parameters of the speech recognition algorithm according to the performance of the word spotting algorithm, the accuracy of the word spotting algorithm can be improved as compared to choosing parameters for the speech recognition algorithm according to an accuracy of that algorithm. For example, if the subword units are phonemes, the parameters of the speech recognition system are not necessarily chosen to optimize phonemic transcription accuracy and can rather be chosen to optimize word spotting accuracy.

[020] Use of spoken queries, as opposed to text-based queries, allows hands-free operation of an audio search engine. For example, in applications such as video editing, an operator may not have hands free to easily use a keyboard, but can speak portions of dialog which is then located.

[021] Queries can be processed without necessarily having a dictionary or letter-to-sound rules for a target language. Furthermore, processing of the query can be optimized for cross-language applications in which the query is spoken in a first language but the second speech recognition system has been trained for a second language.

[022] Other features and advantages of the invention are apparent from the following description, and from the claims.

## Description of Drawings

[023] FIG. 1 is a block diagram of a word spotting system.

[024] FIG. 2 is a looped phoneme grammar.

[025] FIG. 3 is network representation of a query.

[026] FIG. 4 is a network representation of a query formed using an n-best approach.

## Description

[027] Referring to FIG. 1, a word spotting system 100 uses a spoken query 140 to process unknown speech 170 to locate putative query instances 190 associated with the query in the unknown speech. Unknown speech 170 includes acoustically-based data which is derived from an acoustic signal by sampling the waveform and optionally computing signal processing features or statistically based quantities based on the waveform independently of the spoken query.

[028] In different modes of operation of the word spotting system 100, the spoken query 140 can be based on one or more of a number of sources of acoustic input, including being based on an utterance by a user of the system, or on a segment of acoustically-based data derived from an earlier acoustic input. For user-based input, a user of the system speaks the query one or more times, and the system processes the acoustic input of the user's utterances. For example, if the user want to locate instances of a person's name (which the user may not know how to spell, or may originate in a foreign language and therefore may not have a well-defined representation using English units), the user speaks the name into a microphone and the system processes that speech to form the query.

[029] The word spotting system 100 includes a query recognizer 150, which includes an implementation of a speech recognition algorithm and which is used to process acoustically-based data associated with the spoken query. The query recognizer 150 produces a processed query 160. The processed query 160 includes a data representation of the query in terms of subword linguistic units, which in this version of the system are English language phonemes. This representation of the query defines one or more possible sequences of subword units that can each correspond to the query. The data representation of the processed query 160 defines a network representation of the query such that paths through the network each correspond to a possible sequence of subword units.

[030] A word spotting engine 180 then uses the processed query 160 to process the unknown speech 170, which is input to the word spotting system 100. The word spotting engine 180 determines time locations at which the query is likely to have occurred, optionally each associated with a score that characterizes a confidence that the query truly occurred there. These time locations are referred to as "putative" query instances

because it is possible that some of the instances do not truly correspond to the query having been spoken at those times.

[031] Both the query recognizer 150 and the word spotting engine 180 make use of Hidden Markov Model (HMM) technology, which make use of subword models 130 that are trained based on training recordings 110. A training system 120 implements a statistical training procedure to determine observation models and state transition probabilities of the subword models. The subword models 130 include a set of English-language phoneme. In this version of the system, each phoneme is represented as a three-state "left-to-right" model. Other forms of HMMs can alternatively be used for the subwork units.

[032] The word spotting engine uses a probability scoring approach in which a probability of the query event occurring is computed for different times in the unknown speech and putative query instances are reported when the probability exceeds a threshold.

[033] There are alternative versions of the query recognizer 150, or ways of configuring the query recognizer, which produces processed query 160. In one version, the query recognizer 150 recognizes the spoken query using a looped phoneme grammar as shown in FIG. 2. The nodes "aa" through "zh" represent the different phoneme subword units, and "pau" represents a model for a silence or inter-word pause.

[034] The processed query 160 produced by the query recognizer can take the form of a network representation of a phoneme lattice. For example, the network shown in FIG. 3 is a network representation of a phoneme lattice associated with a spoken query of the word "jury." This network is generated by first computing a phoneme lattice for the spoken query, and then representing all or an automatically selected subset of elements of the lattice as a network.

[035] Another way for the query recognizer 150 to produce the processed query 160 is to perform an n-best recognition of the spoken query based on the phoneme grammar shown in FIG. 2. Each of the n-best phoneme recognition results is then used to form one branch of a network with parallel branches. Such a network formed by an n-best approach is shown in FIG. 4, again for the word "jury." The n-best lists can be computed from a phoneme lattice determined from the spoken query.

[036] Another way for the query recognizer 150 to produce the processed query 160 is to generate a confusion network from a phoneme lattice. The confusing network

includes a series of parallel combinations of confusable phonemes that together represent the query.

[037] In another alternative, rather than using the phoneme grammar shown in FIG. 2, an n-gram Markov model can be used to introduce prior sequence probabilities for the subword units into the recognition of the spoken query.

[038] Parameters of the query recognizer 150, which affect the processing of a spoken query 140 to form the processed query 160, are chosen so that the resulting processed query 160 yields accurate results when used by the word spotting engine 180. These parameters do not necessarily correspond to parameters that might be chosen to yield the highest phoneme accuracy if the query recognizer were evaluated in those terms.

[039] The choice of values of the parameters for the query recognizer 150 is determined by using a number of reference queries that are processed by the query recognizer for various settings of the parameter values. The different processed queries which correspond to the different values of the parameters are used by the word spotting engine to process known input speech in which the true locations of query events are known but not used by the word spotting engine. After processing the known speech with the word spotting engine, an overall performance is quantified for the various different choices of parameter values, and the set of parameter values that yields the highest performance is chosen for configuring the query recognizer.

[040] Alternative parameter selection approached can also be used. For example, one class of alternative approaches is based on an iteration in which overall performance is measured or estimated for a set of parameter values at each iteration, and the set of parameter values is updated and performance measured repeatedly until a final set of parameter values are chosen when the iteration converges or is terminated by a predetermined rule (e.g., based on the number of iterations or on the change in the parameter values on successive iterations).

[041] Different measures of overall performance can be used, and in general, each of these different measures corresponds to different best set of values of the parameters. For example, one set of parameter values may be best suited to yield a high detection rate for queries, another for the lowest false alarm rate, and another for the best tradeoff between false alarms and detections.

[042] A number of different parameters of the query recognizer have been found to actually affect, or are expected to affect, the accuracy of the word spotting engine. One such parameter is a phoneme insertion factor, typically represented as a penalty that is introduced affect the length in number of phonemes that are produced by the query recognizer. Reducing the penalty generally increases the length of phoneme sequences produced by the query recognizer in processing a spoken query. In experiments, it has been found that this penalty is best set at a value that typically generates more phonemes than are found in phonetic transcription of the query.

[043] Another parameter is a beamwidth in a Viterbi algorithm HMM search carried out by the query recognizer 150. The beamwidth parameter is a pruning parameter that affects which or how many partial hypotheses are pruned during recognition because they are relatively unlikely as compared to the highest or higher scoring hypotheses. With a larger beamwidth parameter, the network representations in the processed query 160 tend to be "fuller" representing a larger number of different possible phoneme sequences for the query.

[044] Another parameter is the number of recognition hypotheses in an n-best approach. The larger "n" the more alternative phoneme sequences are used to represent the query. Yet another parameter relates to the contribution, or weight, of the phoneme n-gram probabilities during recognition of the spoken query.

[045] In some versions of the word spotting system 100, multiple examples of a query are used as the spoken query 140 that is input to the query recognizer 150. For instance, the user can speak a query multiple times. These multiple spoken queries are then combined by the query recognizer to form a single processed query. Each instance of the spoken query can be associated with a distinct portion of the network, for example, by combining the n-best recognition results for each instance, or combined to form a single overall network.

[046] In another way of using multiple examples of a query, rather than the user repeatedly speaking a query, the user identifies portions of an acoustically-based data of an earlier acoustic signal, for example, identifying portions of a previously recorded waveform. The earlier acoustic signal may, for example, be associated with the unknown speech that is input to the system, or be associated with a training corpus of representative acoustic input. In this way, further instances of the same words or phases can be detected.

[047] The approach of using utterances by the user and instances of the query in previous acoustic input can be combined such that some examples come from the user and other examples of the query come from representative acoustic input.

[048] Note that as introduced above, although the discussion uses the phrase "word spotting" and words as examples of queries, queries can equally be phrases, or large units such as sentences, or can even form complex expressions, such as combinations of phrases with "wildcard" or optional portions.

[049] The subword units do not necessarily have to correspond to linguistic units, or to lingustic units in the language contained in the unknown speech 170 that is to be processed by the word spotting engine. For example, a corpus of subword units trained from English can be used for queries spoken in another language. The subword units can correspond to linguistic units from one or more languages, or come from a universal set. The subword units can also be identified using other techniques and do not necessarily correspond to linguistic units.

[050] Optionally, the different model parameters can be used by the query recognizer and by the word spotting engine. For example, the model parameters used by the query recognizer may be matched to the acoustic conditions in which the user is expected to utter examples of a query, while the model parameters used by the word spotting engine may be matched to the acoustic conditions (e.g., noise level) of the unknown speech.

[051] Processing of the spoken queries and unknown speech can optionally be performed in two stages. In a first stage, the processing makes use of the subword models 130 to derive the acoustically-based input of the spoken query and/or of the unknown speech. For example, probabilities that each of the phonemes occur at different times in the input are computed in the first stage. This computation can occur for the unknown speech prior to the query being defined. Similarly, in the mode of operation in which queries are identified in a recording, this preprocessing can occur prior to identifying the portions of the recording that contain the query. Then, processing of the query to form the processed query 160 and processing of the unknown speech to locate the putative query instances each proceed with a separate second state processing. In one alternative of the first stage processing, phoneme lattice is precomputed and later used to process the query and to detect instances of the query. In another alternative, "forward" and/or "backward" probabilities are precomputed and stored and used by the further processing.

[052] A document titled "Fast-Talk Audio Queries" is attached as an appendix to this specification.

[053] It is to be understood that the foregoing description is intended to illustrate and not to limit the scope of the invention, which is defined by the scope of the appended claims. Other embodiments are within the scope of the following claims.

What is claimed is:

1. A method comprising:

accepting query data from one or more spoken instance of a query;

processing the query data including determining a representation of the query that defines multiple sequences of subword units each representing the query; and

locating putative instances of the query in input data from an audio signal.

2. The method of claim 1 wherein processing the query data includes applying a speech recognition algorithm to the query data.

3. The method of claim 1 wherein the subword units include linguistic units.

4. The method of claim 2 wherein locating the putative instances includes applying a word spotting algorithm configured using the determined representation of the query.

5. The method of claim 4 further comprising selecting parameter values of the speech recognition algorithm for application to the query data according to characteristics of the word spotting algorithm.

6. The method of claim 5 wherein the selecting of the parameter values of the speech recognition algorithm includes optimizing said parameters according to an accuracy of the word spotting algorithm.

7. The method of claim 5 wherein the selecting of the parameter values of the speech recognition algorithm includes selecting values for parameters including one or more of an insertion factor, a recognition search beam width, a recognition grammar factor, and a number of recognition hypotheses.

8. The method of any of claims 1 through 7 wherein determining the representation of the query includes determining a network of the subword units.

9. The method of claim 8 wherein the multiple sequences of subword units correspond to different paths through the network.

10. The method of any of claims 1 through 7 wherein determining the representation of the query includes determining an n-best list of recognition results.

11. The method of claim 10 wherein each of the multiple sequences of subword units corresponds to a different one in the n-best list of recognition results.

12. The method of any of claims 1 through 7 wherein accepting the query data includes accepting audio data representing the spoken utterances of the query spoken by a user, and processing the audio data to form the query data.

13. The method of any of claims 1 through 7 wherein accepting the query data includes accepting selection by a user of portions of stored data from a previously accepted audio signal, and processing the portions of the stored data to form the query data.

14. The method of claim 13 further comprising, prior to accepting the selection by the user, processing the previously accepted audio signal according to a first speech recognition algorithm to produce the stored data.

15. The method of claim 14 wherein the first speech recognition algorithm produces data related to presence of the subword units at different times in the audio signal.

16. The method of claim 14 wherein processing the query data includes applying a second speech recognition algorithm to the query data.

17. A system comprising:

a speech recognizer for processing query data from one or more spoken instances of a query;

a data storage for receiving a data representation of the query from the speech recognizer, the data representation defining multiple sequences of subword units representing the query;

a word spotter configured to use the data representation of the query to locate putative instances of the query in input data from an audio signal.

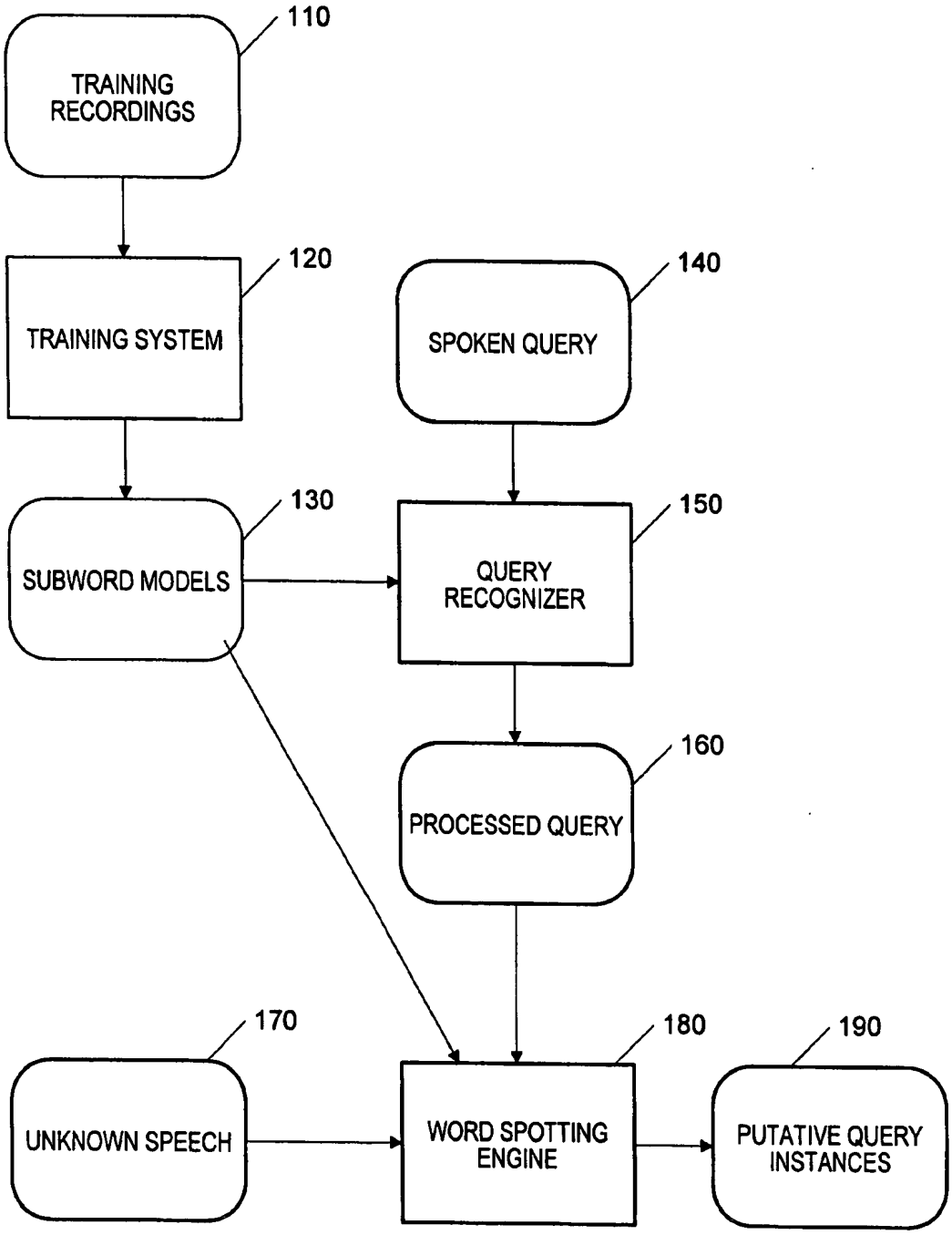
110
TRAINING RECORDINGS
120
TRAINING SYSTEM
140
SPOKEN QUERY
130
SUBWORD MODELS
150
QUERY RECOGNIZER
160
PROCESSED QUERY
170
UNKNOWN SPEECH
180
WORD SPOTTING ENGINE
190
PUTATIVE QUERY INSTANCES

FIG. 1

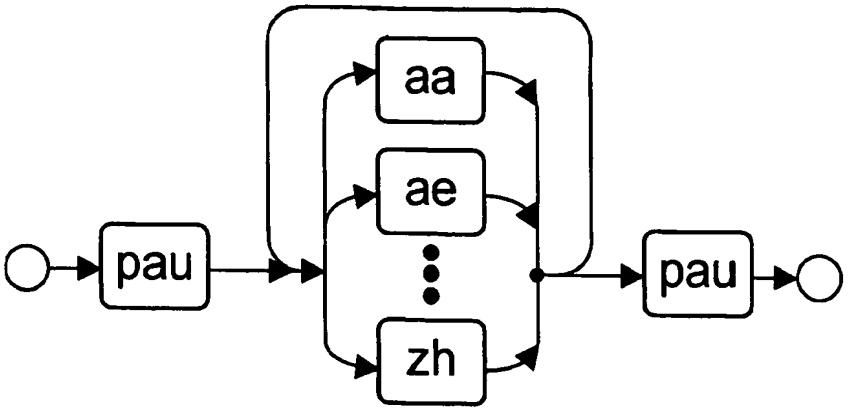


FIG. 2

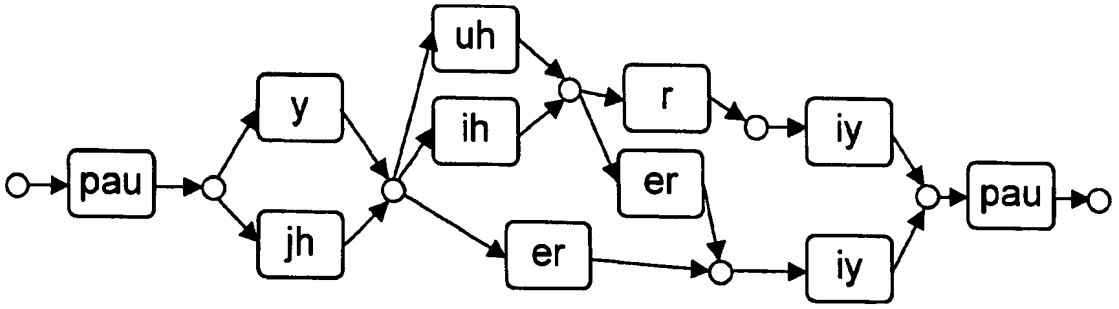


FIG. 3

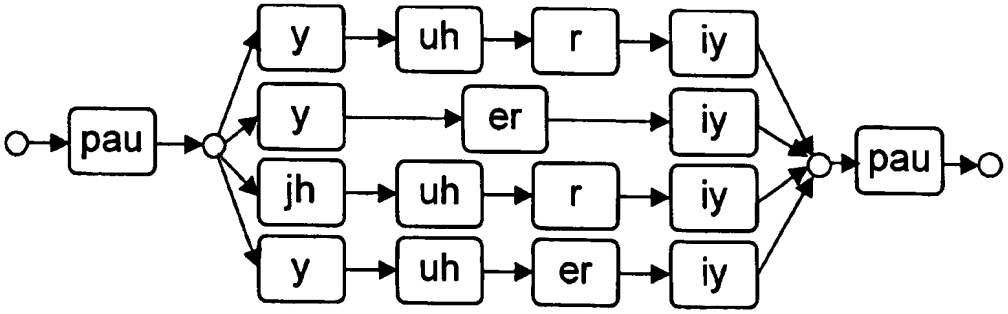


FIG. 4

# This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

## BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☑ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

**IMAGES ARE BEST AVAILABLE COPY.**
**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**